

Method and system for searching URL or Web file and addresses and classifying the search results using an audience indice indicating the frequency of Web address selection

Publication number: FR2802671 (A1)

Publication date: 2001-06-22

Inventor(s): DUTTA RABINDRANATH

Applicant(s): IBM [US]

Classification:


- **international:** **G06F17/30; G06F17/30;** (IPC1-7): G06F17/30


- **European:** G06F17/30W1

Application number: FR20000014661 20001114


Priority number(s): US19990464850 19991216


Also published as:

 FR2802671 (B1)

 US6480837 (B1)

Cited documents:

 WO9948028 (A2)

 US5987457 (A)

Abstract of **FR 2802671 (A1)**

The method uses a server with a Web address which executes the search and generates documents, the results of the file address search including the Web address of the search result indicating the placement of files spread amongst interconnected Internet servers and the generation of the document indicating the search results. Method for classifying searched for data files which has the following stages: (a) association of a key word with each file, a physical location for each file being addressed by a file address; association of an audience indice with at least a pair comprising the file address and key word. A file address can have several associated key words and an audience indice for each file address and key word pair; execution of a search request with search key words; reception of search results for file addresses which have at least an associated key word which corresponds to at least a search key word in response to the search request, and; classification of the results as a function of the audience indice associated with each file address and key word pair in which the key word corresponds the search key word.

20		
21		
22		
www.ibm.com	international	2
..	computer	23
..	software	54
www.uspio.gov	patents	54
..	trademarks	43
..	general	1

.....
Data supplied from the **esp@cenet** database — Worldwide

①⑨ RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE
PARIS

①⑪ N° de publication :
(à n'utiliser que pour les
commandes de reproduction)

2 802 671

②① N° d'enregistrement national : **00 14661**

⑤① Int Cl⁷ : G 06 F 17/30

①②

DEMANDE DE BREVET D'INVENTION

A1

②② Date de dépôt : 14.11.00.

③① Priorité : 16.12.99 US 09464850.

④③ Date de mise à la disposition du public de la
demande : 22.06.01 Bulletin 01/25.

⑤⑥ Liste des documents cités dans le rapport de
recherche préliminaire : *Ce dernier n'a pas été
établi à la date de publication de la demande.*

⑥① Références à d'autres documents nationaux
apparentés :

⑦① Demandeur(s) : INTERNATIONAL BUSINESS MACHI-
NES CORPORATION — US.

⑦② Inventeur(s) : DUTTA RABINDRANATH.

⑦③ Titulaire(s) :

⑦④ Mandataire(s) : IBM FRANCE.

⑤④ METHODE, SYSTEME ET PRODUIT POUR CLASSER DES RESULTATS DE RECHERCHE A L'AIDE D'UN
INDICE D'AUDIENCE.

⑤⑦ La présente invention concerne une méthode, un système, un produit et des structures de données pour classer des fichiers informatiques qui font l'objet d'une recherche. Au moins un mot clé est associé avec chaque fichier. Un emplacement physique de chaque fichier est identifié par une adresse de fichier telle qu'une adresse Web. Un indice d'audience est associé avec au moins une paire constituée d'une adresse de fichier et d'un mot clé, de telle sorte qu'une adresse de fichier peut avoir plusieurs mots clés associés et un indice d'audience associé pour chaque paire constituée d'une adresse de fichier et d'un mot clé. En 5 réponse à l'exécution d'une demande de recherche avec des mots clés, l'utilisateur reçoit des résultats qui ont au moins un mot clé associé qui correspond à au moins un mot clé de la recherche. Les résultats de la recherche sont classés selon l'indice d'audience associé à chaque paire constituée d'un résultat et d'un mot clé dont le mot clé correspond à celui de la recherche. Un document est ensuite codé de manière à comprendre les résultats de la recherche d'adresses de fichier et à les afficher conformément à la classification.

20		
21		
22		
www.ibm.com	international	2
""	computer	23
""	software	54
www.uspto.gov	patents	54
""	trademarks	43
""	general	1

FR 2 802 671 - A1



DOMAINE TECHNIQUE

La présente invention concerne une méthode, un système et un produit pour chercher des adresses de fichier telles que, p. ex., des URLs ou adresses Web et classer les résultats des recherches à l'aide d'un indice d'audience indiquant la fréquence de la sélection des adresses Web parmi les résultats de la recherche renvoyés.

ETAT ACTUEL DE LA TECHNIQUE

Pour trouver des documents sur Internet, les utilisateurs utilisent généralement un moteur de recherche Internet. L'utilisateur entre un ou plusieurs mots clés et indique éventuellement des opérateurs booléens pour la recherche, puis transmet la demande de recherche à un serveur menu d'un moteur de recherche. Les moteurs de recherche comprennent un programme dit "spider" ou "crawler" soit un moteur de balayage qui visite régulièrement les pages web sur Internet pour repérer les nouvelles pages web et vérifier les sites déjà repérés afin de déceler les modifications. Le moteur de balayage place ensuite les informations des pages qu'il a repérées dans un index de base de données qui relie les adresses Web aux termes de recherche.

Les moteurs de recherche peuvent indexer différentes informations des pages repérées pour les associer aux adresses Web repérées. De nombreux moteurs de recherche indexent l'ensemble du texte visible mais peuvent exclure des mots courants tels que, p. ex., "le", "et", etc. Les moteurs de recherche peuvent aussi indexer des mots clés figurant dans un "méta tag" c'est à dire un marqueur méta spécial du document qui comprend des mots clés que le créateur de la page a prévus

pour la recherche. Les moteurs de recherche peuvent inclure d'autres textes associés à des images ou chercher des racines de mots pour inclure des variations telles que, p. ex., politique, politicien, politiquement, etc. comme mots clés dans l'index. Les mots clés indexés pour une adresse Web donnée sont ensuite recherchés lorsqu'un utilisateur les entre pour effectuer une recherche. Le résultat d'une recherche comprend toutes les adresses Web qui possèdent des mots indexés correspondant aux termes de la recherche et aux opérateurs booléens éventuels.

Un moteur de recherche peut donner en réponse à une question d'un utilisateur de nombreux résultats dont beaucoup ne sont pas pertinents. Un des problèmes auxquels les développeurs de moteurs de recherche doivent faire face est l'ordre dans lequel les résultats des recherches doivent être présentés. La plupart des moteurs de recherche utilisent l'emplacement et la fréquence des mots clés sur la page web comme base de classification des résultats de la recherche. D'autres moteurs de recherche peuvent favoriser un ordre d'affichage des pages particulier si des mots clés sont inclus dans les marqueurs méta de la description et des mots clés de la page. Un moteur de recherche peut aussi favoriser la pertinence en fonction du nombre de pages et/ou de pages web importantes qui comprennent des liens hypertexte avec la page des résultats de la recherche.

La réponse à une demande est souvent déterminée par la manière dont les mots clés sont indexés pour une adresse Web. Les fournisseurs de moteurs de recherche contrôlés par les auteurs permettent à ceux-ci de désigner les mots clés indexés en les spécifiant dans les documents. D'autres fournisseurs de moteurs de recherche utilisent une approche de contrôle par les éditeurs; ils emploient alors des éditeurs pour cataloguer manuellement les sites web et les mots clés indexés qui seront utilisés dans la recherche. Les moteurs de recherche qui s'appuient sur les auteurs permettent à ceux-ci d'inclure dans

les marqueurs méta des mots trompeurs pour forcer le moteur de recherche à renvoyer le document en réponse à des recherches sans relation avec son contenu. Les moteurs de recherche contrôlés par les éditeurs peuvent aboutir à des résultats non pertinents si les éditeurs associent des mauvais mots clés aux adresses Web ou excluent des mots clés très pertinents de l'association adresse Web/mot clé.

Pour améliorer la présentation des résultats, une société produisant des moteurs de recherche, Direct Hit, a développé des algorithmes pour classer les résultats de recherche selon l'audience des sites. Le moteur de recherche Direct Hit contrôle anonymement les sites web que les chercheurs sélectionnent dans la liste des résultats, le temps qu'ils passent sur ces sites et d'autres paramètres. Les sites qui sont sélectionnés par les chercheurs sont mieux classés, à l'inverse de ceux qui sont régulièrement ignorés par les chercheurs.

Il y a besoin dans ce domaine d'une méthode, d'un système, d'un programme et de structures de données améliorés pour intégrer l'audience des adresses Web sélectionnées dans l'ordre dans lequel les résultats de recherche sont renvoyés et affichés sur le poste des chercheurs.

RESUME DE L'INVENTION

Pour surmonter les limites de l'état actuel de la technique décrit ci-dessus, les versions préférentielles présentent une méthode, un système, un programme et des structures de données pour classer les fichiers informatiques qui font l'objet d'une recherche. Au moins un mot clé est associé à chaque fichier. L'emplacement physique de chaque fichier est identifié par une adresse de fichier. Un indice d'audience est associé avec au moins une paire constituée d'une adresse de fichier et d'un mot clé, de manière à ce

qu'une adresse de fichier puisse avoir plusieurs mots clés associés et un indice d'audience associé pour chaque paire constituée d'une adresse de fichier et d'un mot clé. En réponse à l'exécution d'une demande de recherche avec des mots
5 clés, l'utilisateur reçoit des résultats qui ont au moins un mot clé associé qui correspond à au moins un mot clé de la recherche. Les résultats de la recherche sont classés selon l'indice d'audience associé à chaque paire constituée d'un résultat et d'un mot clé dont le mot clé correspond à celui de
10 la recherche. Un document est ensuite codé de manière à comprendre les résultats de la recherche d'adresses de fichier et à les afficher selon le classement.

Dans d'autres versions, le système reçoit une demande d'accession à au moins un des résultats de recherche
15 d'adresses de fichier affichés dans le document. Les indices d'audience sont augmentés au moins pour la paire constituée d'un résultat de recherche et d'un mot clé qui correspond aux mots clés de la recherche. Les indices d'audience peuvent aussi être abaissés pour les paires constituées d'une adresse
20 de fichier et d'un mot clé qui correspondent aux mots clés de la recherche qui n'ont pas été demandés.

Dans les versions sur Internet, un serveur doté d'une adresse Web de serveur exécute la recherche et génère le document. Dans ce cas, les résultats de la recherche
25 d'adresses de fichier comprennent des adresses Web qui indiquent l'emplacement de fichiers transmis aux serveurs par Internet. Le document est généré en combinant l'adresse Web de serveur avec chaque adresse Web de résultat de recherche dans une adresse Web combinée pour chaque résultat de
30 recherche. Le document est ensuite codé pour afficher chaque résultat de la recherche avec l'adresse Web combinée de manière à ce que les demandes de l'adresse Web combinée soient orientées vers le serveur. Lorsqu'une demande de l'adresse Web combinée est reçue, l'indice d'audience associé avec

chaque adresse Web de résultat de recherche de la paire constituée de l'adresse Web combinée et du mot clé correspondant aux mots clés de la recherche est augmenté. La demande est dirigée vers l'adresse Web de résultat de la

5 recherche.

Les versions préférentielles comprennent une méthode, un système, un programme et des structures de données pour associer des indices d'audience indiquant la fréquence de sélection, autrement dit l'audience, d'une adresse Web donnée

10 provenant des résultats de la recherche. Dans les versions préférentielles, un indice d'audience peut être associé avec chaque paire constituée d'une adresse Web et d'un mot clé pour donner des indices d'audience spécifiques à chaque mot clé. Les indices d'audience sont utilisés pour déterminer l'ordre

15 dans lequel les résultats de la recherche sont affichés sur le poste du chercheur. Un avantage de la pondération de l'audience des mots clés est que l'ordre de présentation d'une adresse Web donnée dans la réponse à une recherche est basé sur l'audience du mot clé de la recherche et non sur l'adresse

20 Web en général. En fait une adresse Web peut avoir pour différents mots clés des indices d'audience très différents qui indiquent sa pertinence pour les différents mots clés de recherche. Ainsi, les versions préférentielles offrent une classification fine basée sur une audience mesurée qui est

25 spécifique aux mots clés de la recherche et tient compte d'une classification des adresses Web établies par les utilisateurs pour des mots clés donnés.

BREVE DESCRIPTION DES DESSINS

Reportons-nous maintenant aux dessins sur lesquels les

30 numéros représentent toujours les mêmes éléments :

La FIG. 1 représente un environnement informatique dans lequel des versions préférentielles de la présente invention sont installées.

La FIG. 2 représente une exemple de la manière dont les 5 adresses Web, les mots clés et les indices d'audience sont associés dans les versions préférentielles de la présente invention.

Les FIG. 3a et 3b représentent la logique implantée dans un moteur de recherche pour renvoyer les résultats de 10 recherche conformément aux versions préférentielles de la présente invention.

La FIG. 4 représente la logique implantée dans le moteur de recherche pour définir les indices d'audience et traiter les demandes d'adresses Web résultant de la recherche 15 conformément aux versions préférentielles de la présente invention.

DESCRIPTION DETAILLEE DES REALISATIONS DE L'INVENTION

Dans la description suivante, il est fait référence aux dessins d'accompagnement qui en font partie intégrante et qui 20 représentent plusieurs versions de la présente invention. Il est entendu que d'autres versions peuvent être utilisées et des modifications structurelles ou fonctionnelles apportées sans sortir de l'objet de la présente invention.

La FIG. 1 représente un environnement informatique dans 25 lequel des versions préférentielles sont installées. L'environnement informatique 2 comprend un serveur à moteur de recherche 4, au moins un ordinateur client 6 et des serveurs d'information 8a et 8b qui communiquent sur un réseau 10. Le réseau 10 peut être un réseau TCP/IP, tel que l'Internet ou 30 World Wide Web, ou n'importe quel autre système de réseau connu dans le domaine tel que, p. ex., un réseau local (LAN), un Ethernet, un grand réseau (WAN), un réseau système (SAN),

un Token Ring, etc. Les serveurs 4, 8a et 8b et l'ordinateur client 6 peuvent aussi communiquer par des réseaux séparés.

Tous les systèmes 4, 6, 8a et 8b communiquent à l'aide d'un protocole de transfert de document tel que le Hypertext Transfer Protocol (HTTP), ou tout autre protocole de transfert de document connu dans le domaine tel que FTP, Gopher, WAIS, etc. Le client 6 contient un programme d'exploration tel qu'un navigateur HTML, capable de transmettre des demandes à l'aide du protocole HTTP d'une manière connue dans le domaine.

10 Si le client 6 demande l'adresse Web au serveur à moteur de recherche 4, ce dernier renvoie une page HTML dans laquelle l'utilisateur sur le poste client 6 peut entrer des mots clés pour effectuer une recherche sur Internet 10.

Le serveur à moteur de recherche 4 contient un programme de recherche 12 qui effectue des opérations de moteur de recherche connues dans le domaine, telles que la recherche de pages web et l'indexation des adresses Web des pages web avec des mots clés dans un index d'adresse Web 14. Lorsque le client 6 envoie une demande de recherche au serveur à moteur

20 de recherche 4, le moteur de recherche 12 interroge l'index d'adresse Web 14 d'une manière connue dans le domaine pour détecter les adresses Web qui ont des mots clés qui correspondent aux termes de la recherche selon les opérateurs de recherche utilisés, tels que des opérateurs booléens, etc.

25 La FIG. 2 représente un exemple d'index d'adresse Web 14 qui associe une adresse Web (20), un mot clé (21) et un indice d'audience (22), fournissant ainsi un indice d'audience à chaque paire adresse Web/mot clé. L'indice d'audience indique la fréquence de sélection de l'adresse Web associée

30 lorsqu'elle est renvoyée en réponse à une recherche comprenant le mot clé associé à l'adresse Web et à l'indice d'audience.

Les associations adresse Web/mot clé de l'index 14 peuvent être générées à l'aide d'un système d'indexation de moteur de recherche connu dans le domaine pour associer les adresses Web avec les mots clés utilisés pour la recherche.

5 De plus, le moteur de recherche des versions préférentielles peut avoir la capacité d'associer un indice d'audience avec chaque paire adresse Web/mot clé. Les versions préférentielles peuvent ajouter l'association de l'indice d'audience aux index d'adresse Web actuelles utilisées par des

10 moteurs de recherche. Les moteurs de recherche actuels associent les adresses Web avec certains mots clés figurant dans le titre, le document dans son entier, des marqueurs méta de mots clés, etc. Une fois que l'index d'adresse Web est généré d'une manière connue dans le domaine, les versions

15 préférentielles commencent à associer un indice d'audience avec chaque mot clé pour une adresse Web donnée, autrement dit avec chaque paire adresse Web/mot clé. Ainsi, toute adresse Web peut avoir de multiples indices d'audience associés si plusieurs des mots clés associés à l'adresse Web ont fait

20 l'objet de demandes de recherche. En l'occurrence, sur la FIG. 2, les adresses Web "www.ibm.com" et "www.uspto.gov" possèdent de multiples indices d'audience pour chaque paire adresse Web/mot clé. De nombreux moteurs de recherche utilisent tous les mots de la page comme mots clés. Le site

25 de l'Office américain des brevets ayant l'expression "informations générales" sur son portail, le mot clé "général" peut être associé avec l'adresse Web de l'Office des brevets (www.uspto.gov) et avoir un faible indice d'audience car peu de gens effectuant une recherche sur le terme "général"

30 seraient susceptibles d'être intéressés par le site web de l'Office des brevets.

L'indice d'audience peut être associé avec la paire adresse Web/mot clé à l'aide d'une base de données et de techniques d'indexation connues dans le domaine. Dans les

35 versions préférentielles, aucun indice d'audience n'est

associé avec une paire adresse Web/mot clé tant que le mot clé ne fait pas l'objet d'une demande de recherche dans laquelle l'adresse Web correspondant au mot clé est renvoyée dans les résultats. De cette manière, l'indice associé peut être
5 ajouté aux index d'adresse Web courantes. Toutes les paires adresse Web/mot clé peuvent aussi être initialisées avec un indice d'audience qui est augmenté ou diminué selon la fréquence de sa sélection parmi les résultats de la recherche, comme indiqué ci-après.

10 A chaque fois que le moteur de recherche 12 obtient des adresses Web en réponse à une recherche, il génère une page HTML pour afficher les résultats de la recherche classés en fonction de l'indice d'audience de chaque paire adresse Web/mot clé impliquée dans la recherche. Par exemple, si la
15 recherche a porté sur le mot clé "international" et que l'adresse Web "www.ibm.com" a été renvoyée, l'indice d'audience de "www.ibm.com" qui sera utilisée pour déterminer la classification sera "2", c'est à dire l'indice de la paire "www.ibm.com" /"international".

20 Après avoir généré une page HTML pour les résultats de la recherche, le moteur de recherche 12 peut également générer un identifiant unique pour la page HTML, comprenant un champ indiquant le nombre d'adresses Web qui ont été sélectionnées sur la page HTML. Le champ de l'adresse Web indiquant le
25 nombre d'adresses Web déjà sélectionnées sur la page peut être utilisé pour déterminer l'ajustement des indices d'audience de sorte que les adresses Web demandées en dernier aient un niveau de priorité inférieur à celui des résultats de recherche d'adresses de fichier demandés plus tôt.

30 Les FIG. 3a et 3b représentent la logique implantée dans le programme du moteur de recherche 12 pour déterminer les indices d'audience de chaque paire adresse Web/mot clé. Les indices d'audience peuvent être programmés au départ sur une valeur neutre pour pouvoir les diminuer pour indiquer un

niveau de priorité faible pour une paire adresse Web/mot clé donnée. Sur la FIG. 3a, le contrôle commence au bloc 100, lorsque le moteur de recherche 12 reçoit une demande de recherche du client 6. Le moteur de recherche 12 exécute (au 5 bloc 102) la recherche dans l'index d'adresses Web 14 à l'aide d'un des mots clés et des opérateurs éventuels spécifiés dans la recherche. S'il n'y a qu'un résultat de recherche, (réponse "non" au bloc 104), le contrôle passe au bloc 116 sur la FIG. 3b. S'il y a plusieurs résultats de recherche (réponse 10 "oui" au bloc 104), le moteur de recherche détermine alors (au bloc 108) l'indice d'audience pour chaque paire constituée d'un résultat de recherche d'adresse Web et d'un mot clé correspondant au mot clé de la recherche dans l'index d'adresses Web 14. Ainsi, tout résultat de recherche 15 d'adresse Web peut avoir plusieurs indices d'audience si l'index d'adresse Web 14 a plusieurs entrées pour l'adresse Web qui correspondent aux mots clés de la recherche. Par exemple, sur la FIG. 2, une recherche basée sur les mots "computer" et "software" renverrait l'adresse Web IBM avec 20 deux indices d'audience différents car les mots clés de la recherche correspondent à deux mots clés de l'adresse Web IBM.

Si plusieurs indices d'audience correspondent à une adresse Web résultant de la recherche (réponse "oui" au bloc 25 110), le moteur de recherche 12 calcule la moyenne (au bloc 112) des indices d'audience pour cette adresse Web. Ainsi, un seul indice d'audience, ou indice d'audience moyen, est associé avec chaque résultat de recherche d'adresse Web. Les résultats sont ensuite classés (au bloc 114) en fonction de 30 l'indice unique d'audience, ou indice d'audience moyen, de l'adresse Web. D'autres méthodes peuvent être utilisées dans le cas où deux indices d'audience sont associés avec un résultat de recherche d'adresse Web, par exemple en sélectionnant l'indice d'audience maximum de l'adresse Web 35 résultant de la recherche.

Du bloc 114, le contrôle passe au bloc 116 sur la FIG. 3b pour construire une/des page(s) HTML comprenant les résultats de recherche à renvoyer au client 6. Si tous les résultats de recherche peuvent tenir sur une seule page HTML (réponse "oui" 5 au bloc 116), le moteur de recherche 12 établit (au bloc 118) une seule page HTML pour afficher tous les résultats de la recherche classés selon leurs indices d'audience et attribue un identifiant (ID) unique à cette page. Comme nous l'avons indiqué, cet ID unique peut comprendre un champ indiquant le 10 nombre d'adresses Web déjà sélectionnées sur la page, lequel est initialement de zéro. Le moteur de recherche 12 réécrit ensuite (au bloc 120) l'adresse Web correspondant à chaque résultat de recherche en associant l'adresse Web du moteur de recherche et l'identifiant unique de la page. Si tous les 15 résultats de la recherche ne tiennent pas sur une seule page HTML (réponse "non" au bloc 116 aboutissant au bloc 122), le moteur de recherche 12 construit plusieurs pages HTML pour afficher tous les résultats de la recherche classés selon leurs indices d'audience sur chaque page et dans l'ensemble du 20 document. Les résultats de la recherche sont affectés (au bloc 124) à une page et à chaque page est attribué un ID unique qui comprend un champ indiquant le nombre d'adresses Web déjà sélectionnées sur la page. Le moteur de recherche 12 réécrit ensuite (au bloc 126) chaque adresse Web résultant de 25 la recherche en associant l'adresse Web du moteur de recherche et l'ID unique auquel le résultat de la recherche est affecté.

Une fois la/les page(s) établie(s), à partir du bloc 120 ou 126, le contrôle passe au bloc 128 où le moteur de 30 recherche 12 réduit l'indice d'audience de chaque paire adresse Web/mot clé de la première page comme si celle-ci n'avait pas été sélectionnée. La page de recherche est ensuite envoyée (au bloc 130) au client. Ce n'est qu'après la sélection que l'indice d'audience de la paire adresse Web/mot

clé sélectionnée est augmenté. Les indices d'audience des paires adresse Web/mot clé des pages non encore envoyées au client 6 sont abaissés lorsque la page comprenant les résultats de la recherche d'adresse Web est renvoyée au client

5 6. Ainsi, au départ, les indices d'audience de tous les résultats de la page renvoyée à l'utilisateur sont ajustés comme s'ils n'avaient jamais été sélectionnés par l'utilisateur sur le poste client 6. Seuls les adresses Web sélectionnées voient l'indice d'audience de la paire adresse

10 Web/mot clé correspondante augmenté.

La FIG. 4 représente la logique implantée dans le moteur de recherche 12 pour traiter un résultat de recherche d'adresse Web sélectionnée par l'utilisateur sur la page des résultats. L'adresse Web du serveur à moteur de recherche 4

15 étant associée à l'adresse Web résultant de la recherche, c'est l'adresse Web réécrite sélectionnée par l'utilisateur et affichée sur la page HTML renvoyée qui est dirigée vers le serveur à moteur de recherche 4, et non l'adresse Web du site résultant de la recherche sélectionnée par l'utilisateur.

20 Cependant, l'utilisateur est informé de la réorientation car il peut visualiser l'adresse Web du moteur de recherche dans l'adresse Web résultant de la recherche affichée sur la page HTML. Voici un exemple d'adresse Web (1) réécrite associant l'adresse Web du moteur de recherche "www.searchengine.com" à

25 l'adresse Web résultant de la recherche "www.news.com" :

"www.searchengine.com/www.news.com"

(1)

Comme on peut le voir, l'utilisateur du moteur de recherche peut clairement constater que l'adresse Web a été réécrite et

30 que l'adresse Web résultant de la recherche qu'il sélectionne contient un lien avec le site du moteur de recherche. Pour garantir davantage la protection de l'intimité de l'utilisateur et s'assurer qu'il est informé de la réorientation, un message peut être affiché sur la page HTML

35 pour l'informer du fait que la sélection de l'adresse Web va

le diriger vers le serveur à moteur de recherche 4 avant le site sélectionné et que des informations sur les sélections des utilisateurs sont collectées de manière anonyme.

Sur la FIG. 4, le contrôle commence au bloc 150, lorsque
5 le serveur à moteur de recherche 4 reçoit une demande de lecture HTTP de l'adresse Web réécrite du résultat de recherche sélectionné, avec l'ID unique de la page sur laquelle l'utilisateur a sélectionné l'adresse Web. Le moteur de recherche 12 appelle alors (au bloc 152) l'ID unique et
10 augmente la valeur dans le champ de l'ID indiquant le nombre de résultats déjà sélectionnés sur la page. Le moteur de recherche 12 détermine ensuite (au bloc 154) une valeur de pondération à ajouter à l'indice d'audience qui est basée sur le nombre courant de sélection dans le champ de l'ID. Cette
15 augmentation doit aussi tenir compte du fait que l'indice a initialement été réduit (au bloc 128) jusqu'à une valeur par défaut. Le moteur de recherche détermine ensuite (au bloc 156) l'indice d'audience de chaque paire constituée d'un résultat de recherche et d'un mot clé correspondant aux mots
20 clés de la recherche dotés de l'indice déterminé au bloc 154. Le moteur de recherche 12 réoriente ensuite (au bloc 158) la demande HTTP vers l'adresse Web du site sélectionnée où l'utilisateur peut accéder directement au site.

La méthode, le système, le programme et les structures de
25 données des versions préférentielles offrent une technique pour définir l'audience des résultats de recherche afin de déterminer la manière de présenter ou de classer ceux-ci lors des futures recherches. De plus, les versions préférentielles offrent des indices d'audience précis en associant un indice
30 d'audience à chaque mot clé des adresses Web. Seuls les indices d'audience dont les mots clés correspondent aux mots clés de la recherche sont définis. Ainsi, les indices d'audience sont spécifiques à chaque mot clé. Cela est intéressant car l'audience d'une adresse Web peut varier pour
35 différents mots clés de recherche. Par exemple, sur la FIG.

2, les personnes effectuant une recherche avec les mots "general" et "Patton" désirent vraisemblablement visiter des sites web traitant du célèbre général américain George S. Patton. Une recherche utilisant le mot "general" peut
5 renvoyer à l'adresse Web de l'Office américain des brevets ("www.uspto.gov") parce que le mot "general" figure dans l'index d'adresse Web 14. Cependant, les personnes effectuant une recherche avec le mot "general" ne sont vraisemblablement pas intéressées par l'adresse Web de l'Office américain des
10 brevets. C'est pourquoi l'indice d'audience associé à l'adresse Web de l'Office américain des brevets et au mot clé "general" peut être faible. En revanche, l'indice d'audience associé à l'adresse Web de l'Office américain des brevets et au mot clé "patents" sera vraisemblablement élevé car les gens
15 effectuant une recherche sur les brevets seront intéressés par la page web de l'Office américain des brevets.

Ainsi, les versions préférentielles tiennent compte du fait que différentes paires mot clé/adresse Web peuvent avoir différents indices d'audience. Cette précision dans la
20 définition des indices d'audience augmente les chances que la classification des résultats de recherche reflète avec précision la véritable audience des adresses Web pour une recherche donnée, et non des indices d'audience qui ne seraient pas pertinents pour les mots clés utilisés.

25 Les versions préférentielles tiennent aussi compte de l'audience sur une page donnée pour les résultats de recherche qui nécessitent plusieurs pages. Comme nous l'avons indiqué, le nombre de sélections est géré page à page. Ainsi, la première adresse Web sélectionnée sur chaque page de résultats
30 de recherche verra son/ses indice(s) d'audience pour le(s) mot(s) clé(s) augmenté comme s'il était la première adresse Web sélectionnée. Par conséquent, le fait que le moteur de recherche 12 détecte une adresse Web sur une page ultérieure n'a pas d'impact négatif sur l'indice d'audience de l'adresse

Web. Cette caractéristique des versions préférentielles tient compte du fait que les utilisateurs peuvent ne pas lire toutes les pages des résultats de la recherche ou, après avoir trouvé une page qui les intéresse, ne pas tenir compte des pages 5 suivantes. Ainsi, les résultats de recherche des pages suivantes que l'utilisateur n'a pas lues ne sont pas "pénalisés" de figurer sur une page ultérieure. Seuls les indices d'audience des résultats de recherche d'adresses Web qui ont été lues par l'utilisateur sont ajustés.

10 Cette observation conclut la description des versions préférentielles de l'invention. Voici maintenant la description de quelques versions alternatives de la présente invention.

Les versions préférentielles peuvent être mises en œuvre 15 sous forme d'une méthode, d'un appareil ou d'un produit à l'aide de techniques de programmation et/ou de conception ordinaires pour produire des logiciels, des micrologiciels, des matériels ou des combinaisons de ces éléments. Le terme de "produit" (ou de "programme informatique") désigne ici un ou 20 plusieurs programmes informatiques et/ou fichiers de données accessibles à partir d'un ou plusieurs dispositifs ou supports lisibles par ordinateur, tels que des supports magnétiques, des disquettes, des CD-ROM, des disques optiques, des unités holographiques, des mémoires rémanentes, etc. En outre, le 25 produit peut comprendre l'implantation des versions préférentielles sur un support de transmission, tel qu'une ligne de réseau, un support de transmission radio, des signaux se propageant dans l'espace, des ondes radio, des signaux infrarouges, etc. Bien entendu, les spécialistes trouveront 30 de nombreuses modifications qui peuvent être apportées à cette configuration sans sortir de l'objet de la présente invention.

Les versions préférentielles supposaient une disposition particulière de l'index adresse Web pour permettre d'associer

un indice d'audience à chaque paire adresse Web/mot clé. Cependant, les spécialistes des bases de données comprendront que l'alternative avait pour but de permettre l'association des indices d'audience avec les paires adresse Web/mot clé.

5 Sur les versions préférentielles, des indices d'audience sont donnés pour toutes les paires mot clé/adresse Web. En revanche, sur d'autres versions, les indices d'audience peuvent n'être donnés que pour certains mots clés, tels que ceux figurant dans un titre, un marqueur méta de mot clé, etc.

10 Sur les versions préférentielles, l'ordre dans lequel un résultat de recherche d'adresse Web est sélectionné est déterminé en utilisant l'ID unique d'une page avec un champ indiquant le nombre de sélections qui ont été faites sur cette page. Sur d'autres versions, d'autres codes ou structures de
15 données peuvent être utilisés pour déterminer comment définir la priorité des indices d'audience des adresses Web/mots clés pour une page donnée de résultats de recherche.

Les versions préférentielles ont été décrites en envisageant le protocole HTTP pour transmettre les documents
20 entre les ordinateurs d'un réseau. Cependant, les spécialistes comprendront que les versions préférentielles peuvent utiliser n'importe quel protocole de communication pour permettre à un client de demander et d'appeler des fichiers dans un environnement réseau.

25 Les versions préférentielles ont été décrites en envisageant la présentation des résultats de recherche sur des pages HTML. Cependant, on peut utiliser d'autres formats de fichiers pour établir des pages de type web, tels que le Dynamic Hypertext Mark-Up Language (DHTML), le Extensible
30 Markup Language (XML), les Cascading Sytle Sheets, tout autre Standard Generalized Markup Language (SGML) ou autre langage connu dans le domaine pour créer des documents structurés

interchangeables. De plus, on peut utiliser n'importe quelle version de HTML, telle que les versions 2.0, 3.2, 4.0, etc. Sur d'autres versions, le fichier demandé peut se trouver sous n'importe quel autre format - autre qu'un format de type SGML
5 - qui peut être affiché ou exécuté d'une manière quelconque par le client.

Les versions préférentielles ont été décrites en envisageant un environnement réseau dans lequel les pages sont
10 fournies au client à partir d'un serveur sur un réseau tel que l'Internet. Cependant, sur d'autres versions, le serveur à moteur de recherche peut être géré et utilisé par un système informatique unique, tel qu'un ordinateur muni d'une unité de disque dur, pour chercher des fichiers sur le disque dur de
15 l'ordinateur.

Sur les versions préférentielles, les fichiers recherchés sont identifiés par une adresse Web et peuvent se trouver sur différents serveurs communicant par un réseau commun tel que l'Internet. Sur d'autres versions, on peut utiliser n'importe
20 quel principe d'adressage de fichiers, y compris un chemin d'accès de fichier indiquant l'emplacement d'un fichier dans une unité de stockage locale. Dans ce cas, les versions préférentielles peuvent être utilisées pour chercher des fichiers situés dans une unité de stockage locale et non
25 dispersés sur un réseau.

En résumé, la présente invention propose une méthode, un système, un programme et des structures de données pour classer des fichiers informatiques qui font l'objet d'une recherche. Au moins un mot clé est associé avec chaque
30 fichier. Un emplacement physique de chaque fichier est identifié par une adresse de fichier. Un indice d'audience est associé avec au moins une paire constituée d'une adresse de fichier et d'un mot clé de telle sorte qu'une adresse de fichier peut avoir plusieurs mots clés associés et un indice

d'audience associé pour chaque paire constituée d'une adresse de fichier et d'un mot clé. En réponse à l'exécution d'une demande de recherche avec des mots clés, l'utilisateur reçoit des résultats qui ont au moins un mot clé associé qui
5 correspond à au moins un mot clé de la recherche. Les résultats de la recherche sont classés selon l'indice d'audience associé à chaque paire constituée d'un résultat et d'un mot clé dont le mot clé correspond à celui de la recherche. Un document est ensuite codé de manière à
10 comprendre les résultats de la recherche d'adresses de fichier et à les afficher conformément à la classification.

La description ci-dessus des versions préférentielles de l'invention est donnée dans un but d'illustration et de description. Elle ne prétend pas être exhaustive et ne doit
15 pas être interprétée dans un sens limitant l'invention à la forme précise décrite. De nombreuses modifications et variantes sont possibles à la lumière des informations ci-dessus. L'objet de l'invention ne peut pas être limité par cette description détaillée, mais uniquement par les
20 revendications ci-après. Les caractéristiques, exemples et informations ci-dessus constituent une description complète de la fabrication et de l'utilisation des éléments de l'invention. De nombreuses versions de l'invention pouvant être réalisées sans trahir l'esprit ni sortir de l'objet de
25 l'invention, cette dernière réside dans les revendications ci-après.

REVENDEICATIONS

1. Une méthode pour classer des fichiers informatiques qui font l'objet d'une recherche, comprenant les étapes suivantes :

5 association d'au moins un mot clé avec chaque fichier, un emplacement physique de chaque fichier étant identifié par une adresse de fichier,

association d'un indice d'audience avec au moins une paire constituée d'une adresse de fichier et d'un mot clé, une
10 adresse de fichier pouvant avoir plusieurs mots clés associés et un indice d'audience associé pour chaque paire constituée d'une adresse de fichier et d'un mot clé,

exécution d'une demande de recherche avec des mots clés de recherche,

15 réception de résultats de recherche d'adresses de fichier qui ont au moins un mot clé associé qui correspond à au moins un mot clé de recherche en réponse à l'exécution de la demande de recherche et

classification des résultats de recherche en fonction de
20 l'indice d'audience associé avec chaque paire constituée d'un résultat de recherche d'adresses de fichier et d'un mot clé dont le mot clé correspond au mot clé de recherche.

2. La méthode décrite dans la revendication 1, comprenant en outre le codage d'un document indiquant les
25 résultats de la recherche d'adresses de fichier, le document affichant les résultats de recherche d'adresses de fichier en fonction de la classification.

3. La méthode décrite dans la revendication 2, un serveur doté d'une adresse Web de serveur exécutant la
30 recherche et générant le document, les résultats de recherche d'adresses de fichier comprenant des adresse Web de résultat de recherche indiquant l'emplacement des fichiers répartis sur

des serveurs interconnectés par Internet et la génération du document indiquant les résultats de la recherche d'adresses de fichier comprenant les étapes suivantes :

combinaison de l'adresse Web du serveur avec chaque
5 adresse Web de résultat de recherche dans une adresse Web combinée pour chaque résultat de recherche,

codage du document pour afficher chaque résultat de recherche avec l'adresse Web combinée, les demandes adressées à l'adresse Web combinée étant dirigées vers le serveur,

10 réception d'une demande adressée à l'adresse Web combinée,

augmentation de chaque indice d'audience associé avec chaque adresse Web de résultat de recherche dans la paire constituée de l'adresse Web combiné et d'un mot clé

15 correspondant aux mots clés de la recherche et

réorientation de la demande vers l'adresse Web de résultat de recherche.

4. La méthode décrite dans l'une quelconque des revendications 1 à 3, comprenant en outre les étapes suivantes

20 :

détermination de plusieurs indices d'audience associés avec au moins un résultat de recherche d'adresses de fichier pour les résultats de recherche dotés de plusieurs mots clés associés qui correspondent au mot clé de la recherche,

25 calcul de la moyenne des indices d'audience déterminés pour déterminer un indice d'audience moyen pour les résultats de recherche d'adresses de fichier qui ont déterminé plusieurs indices d'audience, les résultats de recherche étant classés en fonction de l'indice d'audience associé avec chaque
30 résultat de recherche doté d'un seul mot clé associé qui correspond aux mots clés de la recherche et de l'indice d'audience moyen pour chaque résultat de recherche doté de plusieurs mots clés associés qui correspondent aux mots clés de la recherche.

5. La méthode décrite dans l'une quelconque des revendications 1 à 4, comprenant en outre les étapes suivantes :

réception d'une demande d'accès à au moins un des
5 résultats de recherche d'adresses de fichier affichés dans le document et

augmentation des indices d'audience pour au moins une
paire demandée constituée d'un résultat de recherche
d'adresses de fichier et d'un mot clé correspondant aux mots
10 clés de la recherche.

6. La méthode décrite dans l'une quelconque des revendications 1 à 5, comprenant en outre la diminution des indices d'audience pour les paires constituées d'une adresse de fichier et d'un mot clé correspondant aux mots clés de la
15 recherche qui n'ont pas été demandés.

7. La méthode décrite dans l'une quelconque des revendications 5 et 6, le réglage de l'indice d'audience en diminution ou en augmentation étant déterminé par le nombre de résultats de recherche du document déjà demandés, les
20 résultats de la recherche d'adresses de fichier demandés en dernier ayant un niveau de priorité inférieur à celui des résultats de la recherche demandés en premier.

8. La méthode décrite dans l'une quelconque des revendications 5 et 6, le réglage de l'indice d'audience en
25 diminution ou en augmentation étant déterminé par le nombre de résultats de la recherche d'adresses de fichier figurant sur un document qui ont déjà été demandés, les résultats de la recherche d'adresses de fichier demandés en dernier ayant un niveau de priorité inférieur à celui des résultats de la
30 recherche demandés en premier.

9. La méthode décrite dans l'une quelconque des revendications 1 à 8, comprenant en outre la génération de

plusieurs documents pour réunir tous les résultats de la recherche d'adresses de fichier lorsque ceux-ci ne peuvent pas tenir dans un seul document, chaque document affichant les résultats de la recherche selon leur classement.

- 5 10. Un système pour classer des fichiers informatiques qui font l'objet d'une recherche, comprenant des moyens pour mettre en oeuvre toutes les étapes de la methode selon l'une quelconque des revendications 1 à 9.
- 10 11. Un produit pour classer des fichiers informatiques qui font l'objet d'une recherche, ledit produit comprenant un support accessible et utilisable par ordinateur et contenant au moins un programme informatique comprenant des instructions adaptées pour permettre à l'ordinateur d'exécuter toute les
- 15 étapes de la méthode selon l'une quelconque des revendications 1 à 9.

FIG. 1

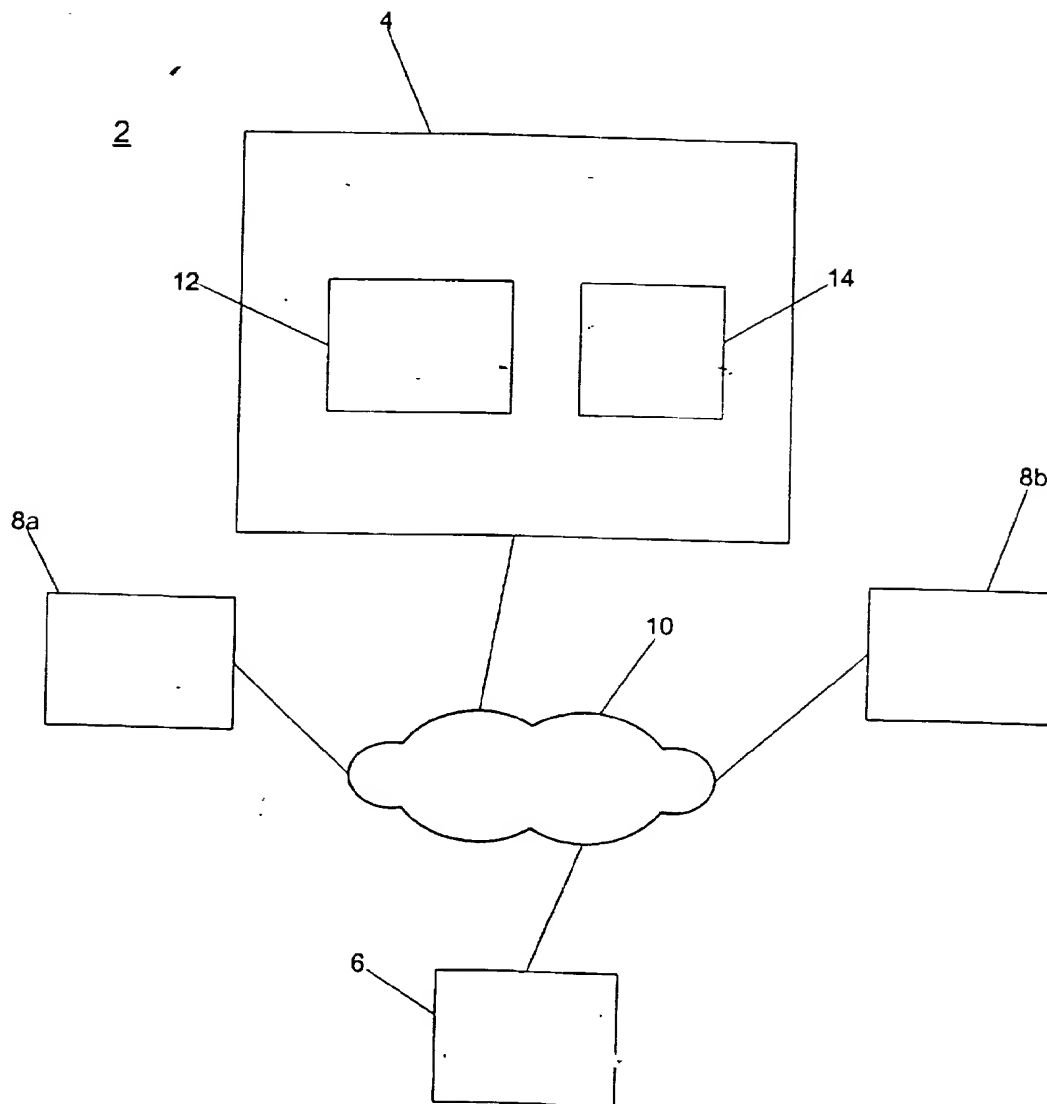


FIG. 2

www.ibm.com	international	2
""	computer	23
""	software	54
www.uspto.gov	patents	54
""	trademarks	43
""	general	1

3/5

FIG. 3a

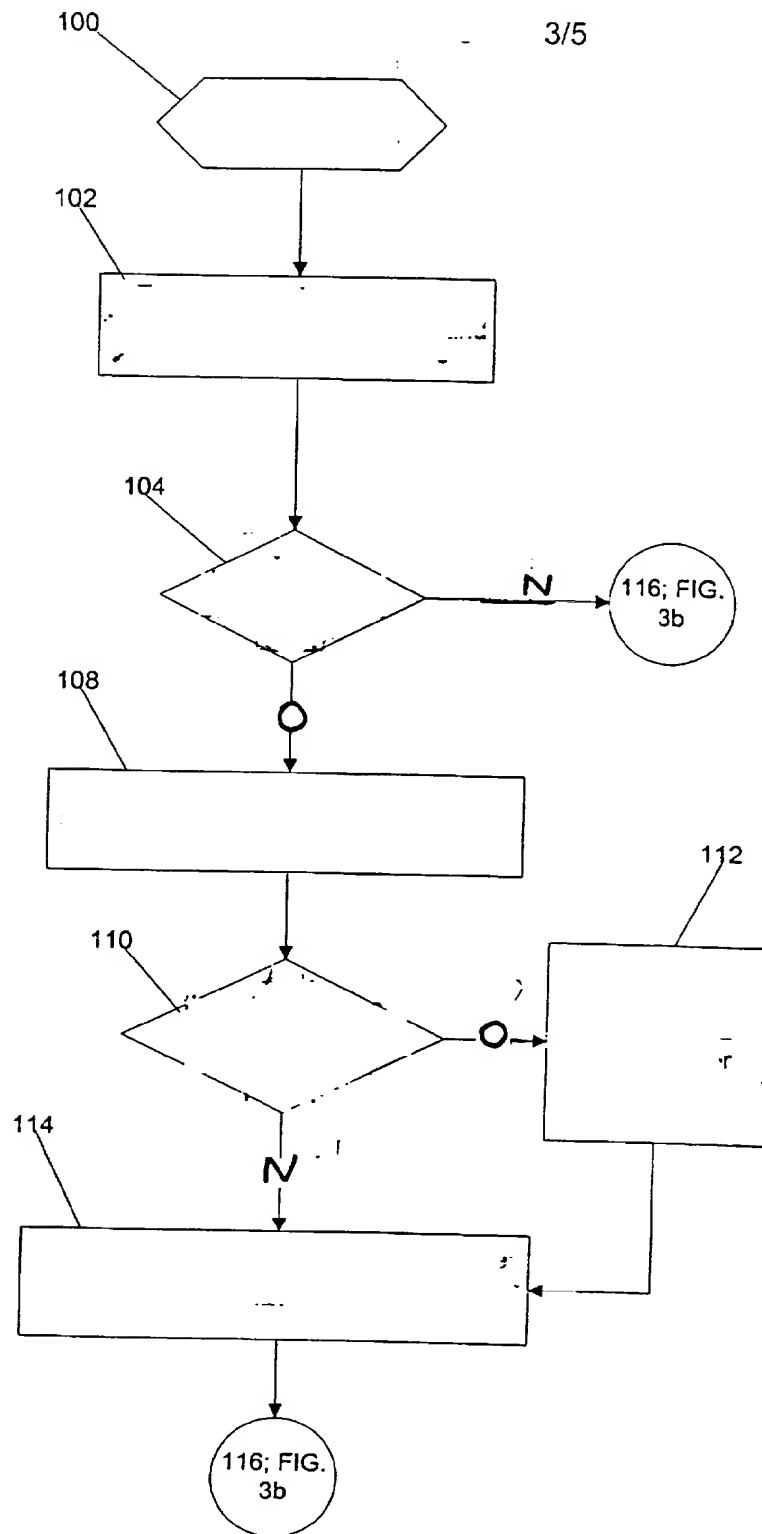
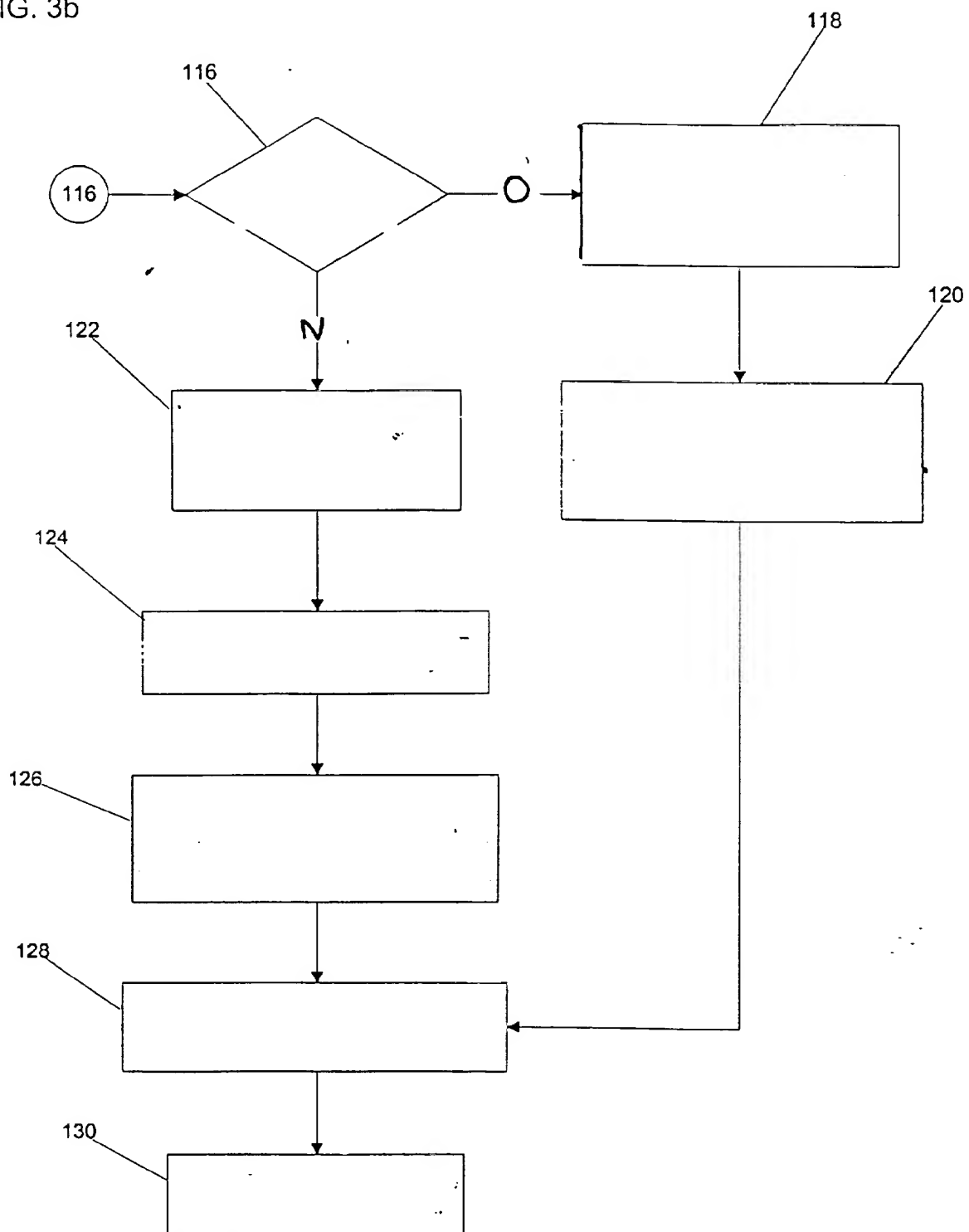


FIG. 3b



5/5

FIG. 4

